Listening. Learning. Leading.®

Research Report
ETS RR–15-11

Effectiveness of Item Response Theory (IRT) Proficiency Estimation Methods Under Adaptive Multistage Testing

Sooyeon Kim

Tim Moses

Hanwook Henry Yoo

June 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Effectiveness of Item Response Theory (IRT) Proficiency Estimation Methods Under Adaptive Multistage Testing

Sooyeon Kim,[1] Tim Moses,[2] & Hanwook Henry Yoo[1]

1 Educational Testing Service, Princeton, NJ
2 The College Board, New York, NY

The purpose of this inquiry was to investigate the effectiveness of item response theory (IRT) proficiency estimators in terms of estimation bias and error under multistage testing (MST). We chose a 2-stage MST design in which 1 adaptation to the examinees' ability levels takes place. It includes 4 modules (1 at Stage 1, 3 at Stage 2) and 3 paths (low, middle, and high). When creating 2-stage MST panels (i.e., forms), we manipulated 2 assembly conditions in each module, such as difficulty level and module length, to see if any interaction existed between IRT estimation methods and MST panel designs. For each panel, we compared the accuracy of examinees' proficiency levels derived from 7 IRT proficiency estimators. We found that the choice of Bayesian (prior) and non-Bayesian (no prior) estimators was of more practical significance than the choice of number-correct versus item-pattern scoring. For the extreme proficiency levels, the decrease in standard error compensated for the increase in bias in the Bayesian estimates, resulting in smaller total error. Possible score changes caused by the use of different proficiency estimators would be nonnegligible, particularly for the extreme proficiency level examinees. The impact of misrouting at Stage 1 was minimal under the MST design used in this study.

Keywords Multistage testing; proficiency estimator; item response theory; routing accuracy

doi:10.1002/ets2.12057

Computerized adaptive tests (CAT) and multistage tests (MST) are both adaptive. However, the testing designs differ substantially in terms of their adaptive algorithm. Under the CAT design, adapting to an examinee's ability occurs at the item level to improve precision and efficiency of measurement, whereas under the MST design, adapting to an examinee's ability occurs between item sets (modules) based on cumulative performance on previous item sets. Examinees receive preassembled item sets determined by their performances at previous stages. MST provides a compromise between fully adaptive testing (e.g., CAT) and nonadaptive testing (e.g., conventional linear forms). This feature has led to interest in MST (see Luecht & Sireci, 2011) and its operational use in practice (see Educational Testing Service [ETS], 2011).

An MST design consists of a small number of separate modules, and each module can be assembled to meet a set of specifications such as item content and item difficulty. The choice of MST design configurations and psychometric characteristics of MST assembly is influenced by various factors, such as test score use (certification or noncertification), item security, item pool capacity, and administration environments. Adding stages and modules within stages can produce tremendous practical complexity without adding psychometric benefits for the final forms (Jodoin, Zenisky, & Hambleton, 2006; Luecht & Nungester, 1998; Luecht, Nungester, & Hadidi, 1996; Wang, Fluegge, & Luecht, 2012). A recent study (Wang et al., 2012) showed that both complex and simple MST designs perform equally well with an optimal item bank consisting of high-quality items targeting key ability regions.

Figure 1 illustrates an example of a two-stage MST panel (i.e., form) in which one adaptation to the examinees' ability levels takes place. Stage 1 (often called routing), uses only one module; all examinees taking that panel are tested with same set of items. At Stage 2, three modules are used: a low-difficulty module, a medium-difficulty module, and a high-difficulty module. Each module at Stage 2 concentrates on a particular level of difficulty to differentiate examinees' abilities within a certain range of proficiency after routing. The items an examinee receives at Stage 2 are determined by the examinee's performance on Stage 1. The term *path* can be used to mean a combination of modules that could possibly be presented to an examinee. In the example of Figure 1, each path consists of the first-stage module and one of the second-stage modules. Three paths in the two-stage MST panel are illustrated in Figure 1. Many variations are possible within the same

Corresponding author: S. Kim, E-mail: skim@ets.org

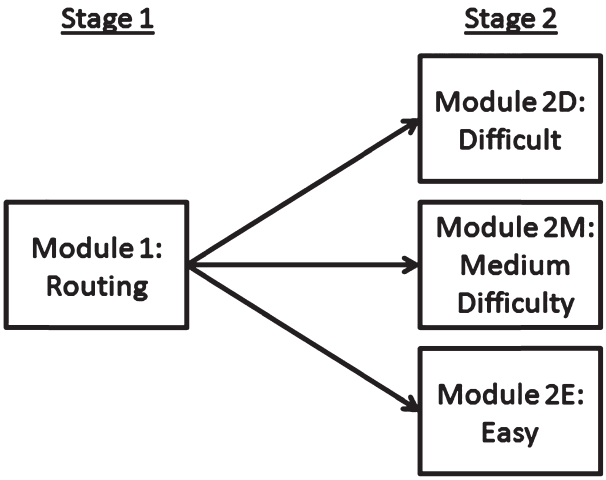

Figure 1 Schematic of a two-stage multistage test (MST).

two-stage MST design as a function of module length. The determination of how to allocate items to the routing section and subsequent sections depends on the test's purpose and the item pool's quality.

Proficiency Estimation Methods

An examinee's proficiency level can be estimated in a variety of ways when using an item response theory (IRT) model, and an MST panel (i.e., form) can be scored in several ways. Generally, well-known IRT proficiency estimation methods are as follows:

1. Test characteristic function with number-correct scoring[1] (TCF)
2. Maximum likelihood estimation with item-pattern scoring (MLE)
3. Expected a posteriori with number-correct scoring ($_S$EAP)
4. Expected a posteriori with item-pattern scoring (EAP)
5. Maximum (mode) a posteriori with item-pattern scoring (MAP).

MLE has been used commonly in practice. MLE finds the examinee's proficiency level that maximizes the likelihood of obtaining the examinee's observed test data, given the item parameters and model. In other words, MLEs are the parameter values that maximize the likelihood that the observed data would have been generated. Thus MLE values correspond to the mode of the likelihood function.[2] Desirable properties of MLE are that it is asymptotically unbiased and that its standard error is related to the information function (Baker, 1992). Drawbacks of MLE, however, include infinite estimates for examinees whose response patterns are either only incorrect (extremely low proficiency) or only correct (extremely high proficiency). Bayesian methods such as EAP and MAP do not share this limitation. Under the Bayesian paradigm, the posterior distribution of the proficiency levels (i.e., θ) is defined as the product of the likelihood function and the prior ability distribution. Bayesian methods incorporate information about the prior distribution to approximate the posterior distribution of latent proficiency (Bock & Mislevy, 1982). The mean of the posterior distribution is the proficiency estimate under EAP, whereas the mode of the posterior distribution is the proficiency estimate under MAP (Yen & Fitzpatrick, 2006). The choice of a reasonable prior distribution for the proficiency level is key to Bayesian estimators. The most common prior distribution is the standard normal distribution, $N(0, 1)$, and the estimated proficiency levels are shrunk toward the prior mean value. Bayesian estimators generally lead to biased estimates, but their overall errors tend to be relatively small due to shrinking to the mean (EAP) or mode (MAP). The shrinkage is expected to be more pronounced under number-correct scoring, owing to its lower precision, than under item-pattern scoring (Kolen & Tong, 2010).

The practical benefits of number-correct scoring versus the psychometric benefits of item-pattern scoring have been debated. TCF (often called *inverse test characteristic curve* [TCC]) uses number-correct scoring (see Stocking & Lord, 1983), and MLE uses item-pattern scoring. Examinees with the same number-correct score but different item responses are likely to obtain different proficiency estimates under item-pattern scoring but the same estimates under number-correct scoring. Number-correct scoring is easier for test users to understand than item-pattern scoring. However, certain

testing programs prefer item-pattern scoring to number-correct scoring, because item-pattern scoring offers more precise estimation than does number-correct scoring. For the three-parameter logistic (3PL) IRT model, not only are standard errors based on item-pattern scoring theoretically lower than standard errors for number-correct scoring, but also empirical evaluations of the standard errors have verified the model predictions for a wide range of multiple-choice (MC) tests (Green & Yen, 1983; Kolen & Tong, 2010; Yen, 1984; Yen & Candell, 1991). According to those studies, item-pattern and number-correct proficiency estimates are equal on average but differ in their standard errors. In general, the longer the test and the less the relative influence of guessing are (e.g., for MC items with many effective distracters or selected responses), the more similar the results from the two scoring methods will be. According to Yen and Fitzpatrick (2006), when tests include "30 or more items, the inverse of the TCC provides a very accurate MLE of ability for the 3PL model" (p. 137). Thus, the distinction between the two scoring methods may not always be clear.

In particular, Yen (1984) showed an interesting idea that an MLE proficiency estimate can be obtained given the examinee's number-correct score or other weighted raw score. In this approach, the MLE for the proficiency estimate is the value that assigns the highest probability to the examinee's number-correct score. Yen used a Taylor series approximation to the compound binomial probability distribution to estimate the proficiency that maximizes the likelihood of obtaining the observed number of correct items. It was somewhat questionable, however, how to achieve the maximum likelihood using such an approximation approach. We think that the Lord and Wingersky (1984) recursion algorithm can be considered a better approach in that this algorithm is capable of computing almost exactly the compound binomial probability distribution for the number-correct score.

A few researchers compared the performance of IRT proficiency estimators and their impact using either real or simulated data sets (Magis, Beland, & Raiche, 2011; Tong & Kolen, 2007, 2010). In a vertical scaling context, Tong and Kolen (2007) showed that different estimators produced score distributions with different characteristics, although they did not affect the growth interpretations from grade to grade. Using real data examples, Kolen and Tong (2010) demonstrated that the choice of estimators (MLE, TCF, EAP, and $_S$EAP) can have a significant influence on practical applications of IRT, such as the score distribution or the assignment of examinees to proficiency levels. They manipulated a test including 53 MC items to create tests that differed in terms of length and difficulty to compare estimation results from the various testing contexts. In their comparisons, the estimation results were somewhat similar when the test was long and included items that varied in difficulty; however, performance classification levels were inconsistent among the estimators. Differences in score distributions for MLE and EAP estimators became more salient as tests became shorter and less reliable. Because the testing conditions were manipulated using the real data set, however, no clear criterion was present with which to compare each estimator's performance. Kolen and Tong (2010) recommended that simulation studies be used so that better conclusions can be drawn regarding the various estimators' accuracy.

Purpose

With IRT, examinees' proficiency can be estimated using maximum likelihood, TCF, or Bayesian estimators. The purpose of this study is to determine which IRT proficiency estimators are the most effective for discovering examinees' true proficiency levels under various MST panel assembly conditions. We chose a two-stage MST design that includes four modules (one at Stage 1 and three at Stage 2) and three paths (low, middle, and high). We assembled eight two-stage MST panels by manipulating two assembly conditions, difficulty level and module length, in each module. For each MST panel, we assessed the two-parameter logistic (2PL) IRT proficiency estimation methods' performance in terms of systematic estimation error (i.e., bias) and random estimation error as a function of MST panel assembly conditions.[3] We compared psychometric effectiveness of number-correct scoring versus item-pattern scoring under MST and examined the usefulness of prior information in enhancing overall precision of proficiency estimates. We also examined the effect of proficiency estimators on the assignment of examinees to the second-stage module to assess the estimators' routing accuracy.

Methods

Multistage Testing Panel Assembly

A total of eight multistage testing (MST) panels (forms) were assembled based on two assembly conditions. One condition was the difference in difficulty among the three second-stage modules, and the other was module length at Stage 1 and Stage 2. Specific levels of each condition were as follows:

1. Difficulty difference among the second-stage modules (two levels)
    a. Small difference in difficulty (i.e., overlap in difficulty)
    b. Large difference in difficulty (i.e., distinction in difficulty).
2. Module length at Stage 1 and Stage 2 (four levels)
    a. More items in Stage 1: Stage 1 = 25 items; Stage 2 = 15 items per module (25 – 15 – 15 – 15)
    b. Equal number of items in each module: 20 items per module (20 – 20 – 20 – 20)
    c. More items in Stage 2: Stage 1 = 15 items; Stage 2 = 25 items per module (15 – 25 – 25 – 25)
    d. Many more items in Stage 2: Stage 1 = 10 items; Stage 2 = 30 items per module (10 – 30 – 30 – 30).

Eight MST forms were created to represent each of the eight assembly conditions based solely on statistical specification (e.g., item difficulty, item discrimination, and number of items). To make the simulated panels realistic, we examined the statistical properties of more than 1,000 MST panels that had been administered in actual operational settings. Under the small difference in difficulty condition, the average of item difficulty parameters was set to be 0.00 for routing, −0.75 for low, 0.00 for middle, and +0.75 for high. The other four MST panels under the large difference in difficulty condition were simply adapted from the small-difference panels. For example, to create large-difficulty difference panels, we subtracted 0.5 from the item difficulty parameter values in the low module of the small-difference panels and added 0.5 to the parameter values in the high module. The average item difficulty parameters of the Stage 1 and Stage 2 middle modules were set at 0.00 in both small- and large-difference conditions. The average of discrimination parameters was set at 0.85 for all 32 modules (2 difficulty levels × 4 module lengths × 4 modules per MST panel). The standard deviation of the item discrimination parameters was 0.27 – 0.30 for each module, resulting in parameters ranging from 0.22 to 1.45.

Item parameters for each of the four modules on a particular MST panel were generated using the Microsoft Excel random number generator function. Figure 2 graphically presents the difficulty differences between the two assembly conditions using box-and-whisker plots. We used each panel's TCC and test information function (TIF) to assess the extent to which each module and each path were reasonable. As an example, Figure 3 presents the TIF of each module as well as each path and the TCC of each path under the 20 – 20 – 20 – 20 module-length condition. Three plots in the first column indicate the TIF and TCC under the small-difference condition, and plots in the second column indicate TIF and TCC under the large-difference condition. Perhaps the TIF is more useful under item-pattern scoring, whereas the TCC is more useful under number-correct scoring in an actual assembly setting. Because we used both scoring methods in this simulation, we considered both the TIF and the TCC in assessing the reasonableness of the simulated MST panels.

Proficiency Estimators

Table 1 summarizes seven IRT proficiency estimation methods that were investigated in this study. Along with the five well-known IRT estimation methods (TCF, MLE, $_S$EAP, EAP, and MAP), we also included two methods that find the proficiency estimate that maximizes the likelihood of obtaining the observed number of correct items. One is Yen's (1984) maximum likelihood estimation ($_Y$MLE), and another is a revised version of $_Y$MLE that applies the Lord and Wingersky (1984) recursion algorithm (i.e., MLE with summed scoring [$_S$MLE]).

TCF, $_Y$MLE, $_S$MLE, and MLE are non-Bayesian estimators, whereas $_S$EAP, EAP, and MAP are Bayesian estimators. TCF, $_Y$MLE, $_S$MLE, and $_S$EAP use number-correct scoring, whereas MLE, EAP, and MAP use item-pattern scoring. The specific formulas for each estimation method are as follows.

Estimates of the examinees' proficiency levels ($\widehat{\theta}$) were obtained by treating the item difficulties (b) and discriminations (a) as known parameters in the H items' 2PL IRT models (Equation 1) and using each of the seven proficiency estimators:

$$P_h\left(a_h, b_h, \theta\right) = \frac{1}{1 + \exp\left[-1.702a_h\left(\theta - b_h\right)\right]}. \quad (1)$$

Using TCF, examinees' proficiency estimates were defined as those that produced summed scores for the 2PL IRT model (i.e., true scores under the IRT model) that closely approximated examinees' summed total test scores, such that

$$\sum_h u_h \approx \sum_h P_h\left(a_h, b_h, \theta\right) \text{ or } X \approx \tau\left(\theta\right), \quad (2)$$

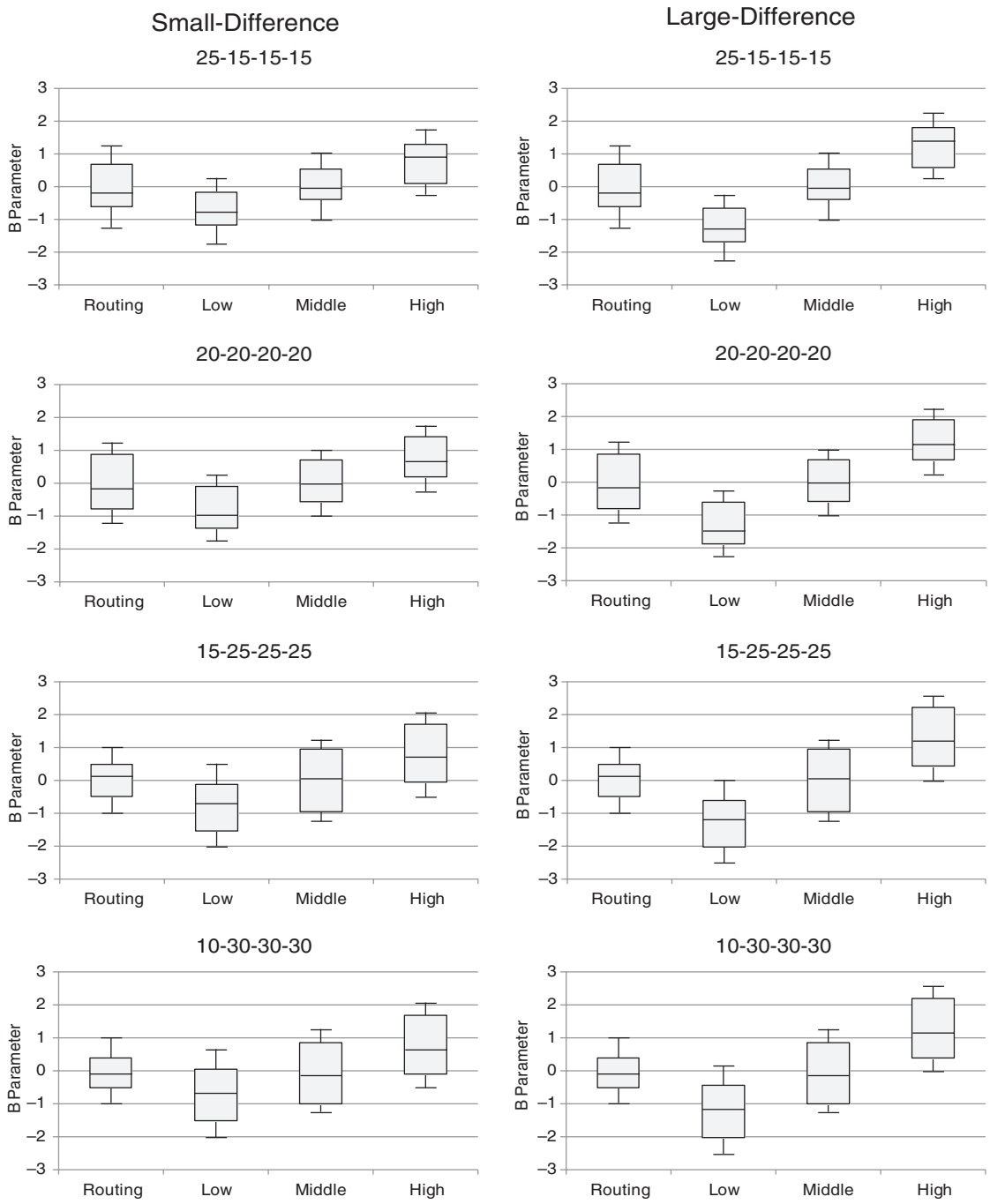

Figure 2 Box-and-whisker plots for each of four modules under conditions of small- and large-difficulty differences.

where u_h indicates the sum of correct responses to the H items, X indicates the summed score, and τ indicates true summed score (Kolen & Tong, 2010). Using MLE, examinees' proficiency estimates were defined as the θ values that maximized the likelihood of examinees' observed patterns of responses to the H items ($U = (u_1, u_2, \ldots, u_H)^t$), as shown in Equation 3:

$$L(U, \theta) = \prod_h \left\{ P_h\left(a_h, b_h, \theta\right)^{u_h} \left[1 - P_h\left(a_h, b_h, \theta\right)\right]^{(1-u_h)} \right\}. \quad (3)$$

The maximization was accomplished using the Newton–Raphson algorithm (Baker & Kim, 2004).[4] As shown in Equation 4, examinees' proficiency estimates can be also defined as the proficiency values based on the likelihood of

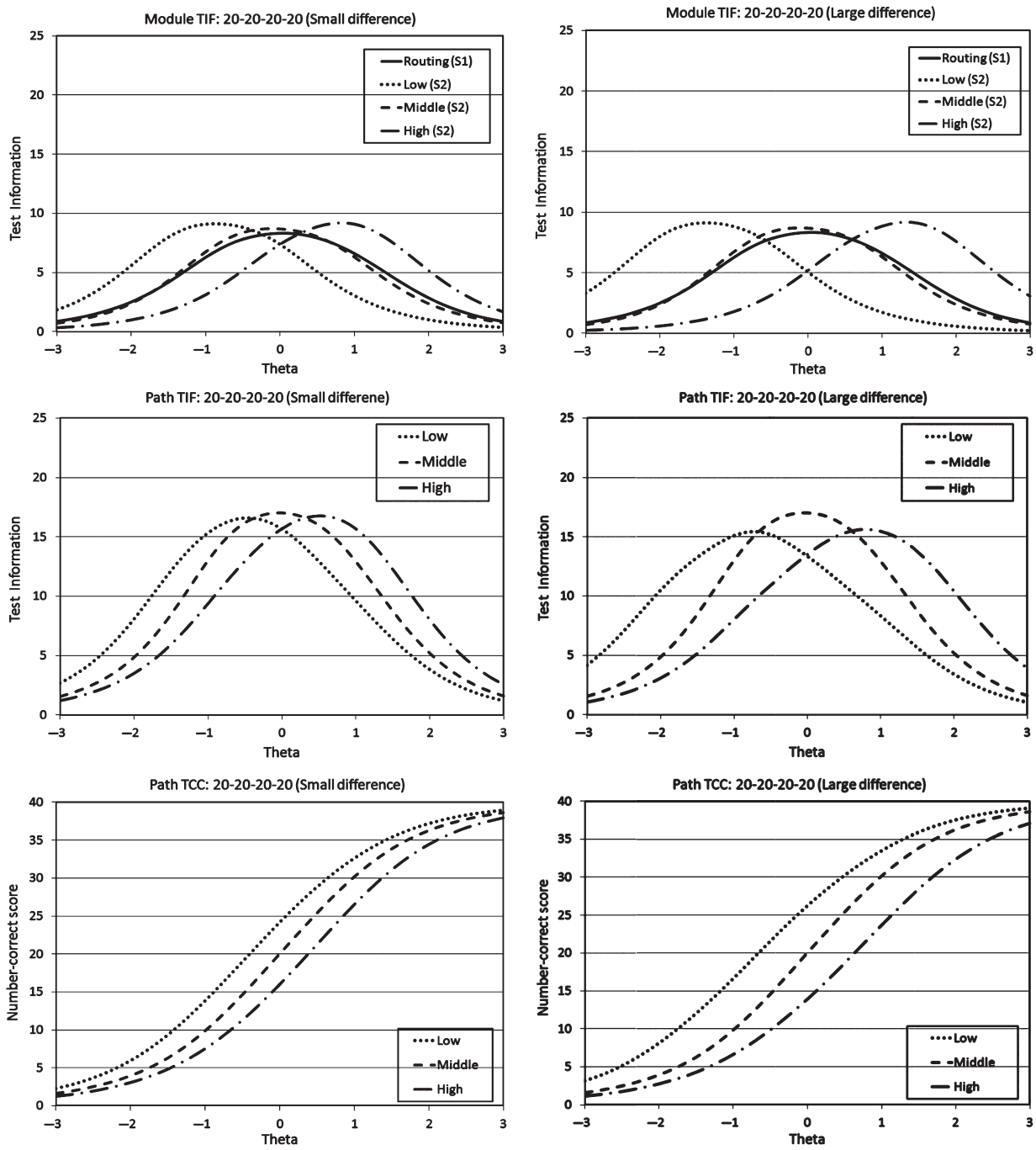

Figure 3 Test information function (TIF) and test characteristic curves (TCC) for each module and path under the small- and large-difference in difficulty conditions of the 20 – 20 – 20 – 20 panel.

their summed total test scores ($X = \sum_h u_h$):

$$P(X|\theta) = \sum_{\substack{\text{response} \\ \text{patterns} \\ \text{with} \sum_h u_h = X}} \prod_h \left\{ P_h\left(a_h, b_h, \theta\right)^{u_h} \left[1 - P_h\left(a_h, b_h, \theta\right)\right]^{(1-u_h)} \right\}. \quad (4)$$

Table 1 Characteristics of the Item Response Theory (IRT) Proficiency Estimation Methods

	Scoring	
Prior distribution	Number correct	Item pattern
No (non-Bayesian)	TCF, $_{Y}$MLE, $_{S}$MLE	MLE
Yes (Bayesian)	$_{S}$EAP	EAP, MAP

Note. TCF = test characteristic function; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$MLE = maximum likelihood estimation with summed scoring; MLE = maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori; MAP = maximum (mode) a posterior.

Using $_{Y}$MLE, $P(X|\theta)$ can be computed using a Taylor series approximation to the compound binomial probability distribution of the number-correct score. Using $_{S}$MLE, $P(X|\theta)$ can be computed using the Lord and Wingersky (1984) recursion algorithm as a method to compute the distribution of the number-correct score. Using $_{S}$EAP, examinees' proficiency estimates were defined as average proficiency values based on the likelihood of their summed total test scores ($X = \sum_h u_h$) and an assumed standard normal proficiency distribution:

$$_{S}EAP = E(\theta|X) = \frac{\int_{\theta} \theta L(X,\theta)\, g(\theta)\, d\theta}{\int_{\theta} L(X,\theta)\, g(\theta)\, d\theta}, \tag{5}$$

where $L(X,\theta)$ was computed using the recursion algorithm and the assumed $g(\theta)$ distribution was approximated with a discrete distribution containing 41 θ values and quadrature points. Using EAP, examinees' proficiency estimates were defined as average values based on the likelihood of their response patterns and an assumed standard normal proficiency distribution:

$$\text{EAP} = E(\theta|U) = \frac{\int_{\theta} \theta L(U,\theta)\, g(\theta)\, d\theta}{\int_{\theta} L(U,\theta)\, g(\theta)\, d\theta}. \tag{6}$$

As in operational practice, the standard normal distribution for θ and the integration of this distribution were accomplished by approximating the continuous proficiency distribution, $g(\theta)$, using a discrete distribution with 41 θ values and quadrature points (Baker & Kim, 2004). EAP and MAP follow essentially the same equation. Using MAP, however, examinees' proficiency estimates were defined as those that maximized the likelihood of examinees' response patterns for an assumed standard normal proficiency distribution, $L(U,\theta)g(\theta)$, as appearing in Equation 6. Finding the maximum value was accomplished using the Newton – Raphson algorithm (Baker & Kim, 2004).

Procedures

We simulated data from 2,000 examinees at each of 41 quadrature points on a theta scale ranging from −3.0 (minimum) to +3.0 (maximum), with an interval of 0.15. Accordingly, simulated examinees' thetas were uniformly distributed ($N = 82{,}000$). In the present study, the framework is restricted to conditional estimation of proficiency levels, given that the item parameters are known in advance. With fixed item parameters, the proficiency levels can be estimated using several methods. The simulation procedure for each examinee, for each of the eight MST panels, consisted of the following steps:

1. Generate the simulated examinee's response to each item in the Stage 1 module.
2. Compute the examinee's proficiency (i.e., theta) on Stage 1 using each of the seven proficiency estimators and assign the examinee to the appropriate Stage 2 module according to the provisional proficiency.
3. Generate the examinee's response to each item in the Stage 2 module.
4. Compute the examinee's proficiency on Stage 1 and Stage 2 combined, using each of the seven proficiency estimators, to determine the examinee's estimated proficiency theta.

We replicated this procedure for 82,000 examinees, each using eight panels of the MST. We used SAS statistical software to generate simulated examinees' dichotomous responses to each item of the MST panel and to estimate the proficiency levels.

Luecht, Brumfield, and Breithaupt (2006) discussed two methods for locating the routing points on the proficiency (θ) scale. One is the *approximate maximum information* method, which finds the optimal decision point on the θ scale for selecting between two modules, using a maximum information criterion similar to any computerized adaptive test. The other is the *defined population intervals* method, which can be used to implement a policy that specifies the relative proportions of examinees in the population expected to follow each of the available paths through the panel. Currently, at least one large-scale international testing program uses the defined population intervals method to determine cutscores, and we used the same method in this simulation. We selected two cutscores for routing simulated examinees to a target module so as to result in approximately 30%, 40%, and 30% of the examinees taking high, middle, and low paths, respectively, in a situation where the simulated examinees' ability distribution follows the standard normal distribution — that is, $\theta = N(0, 1)$. Given this condition, the proficiency cutscores associated with the 30th and 70th percentiles of the cumulative distribution of thetas would be approximately −0.5 and +0.5. The cutscores were essentially the same in both small- and large-difference conditions because their routing modules were identical.

Given that all simulated examinees' true proficiency levels (i.e., thetas) are known, their estimated thetas can be compared to their true thetas to evaluate the effectiveness of the seven estimators. On the basis of the difference between estimated and true values, three deviance measures, bias, error, and root mean squared error (RMSE), were calculated at each of the 41 quadrature points using the following formulas:

$$\text{Bias}_{ij} = \left(\widehat{\theta}_{ij} - \theta_i\right), \tag{7}$$

$$\text{Error}_{ij} = SD\left(\widehat{\theta}_{ij}\right), \tag{8}$$

$$\text{RMSE}_{ij} = \sqrt{\text{Bias}_{ij}^2 + \text{Error}_{ij}^2}, \tag{9}$$

where i indicates a theta point, j indicates a proficiency estimator, $\widehat{\theta}_{ij}$ indicates a proficiency estimate of a particular estimator, and θ_i indicates a true proficiency value. As overall summary measures, root mean squared bias, error, and RMSE were each averaged across all quadrature points, weighting the values at each theta level according to its relative percentage (f_i) under the standard normal distribution.[5] The resulting statistics were the weighted root mean squared bias, $\sqrt{\sum_i f_i \text{Bias}_{ij}^2}$, the weighted standard error of estimation, $\sqrt{\sum_i f_i \text{Error}_{ij}^2}$, and the weighted RMSE, $\sqrt{\sum_i f_i \text{RMSE}_{ij}^2}$. We used a root mean square averaging procedure to prevent negative bias at one score level from canceling out positive bias at another. Percentage classification accuracy at Stage 1 was also calculated for each of the proficiency estimators to determine the extent to which the proficiency estimators performed differently as a function of the number of items used for routing.

Results

For each of the seven proficiency estimators, we compared simulated examinees' estimated proficiency levels (i.e., theta values) to their true proficiency levels. In Figure 4, the four plots in the first column present estimation bias derived using the 25 – 15 – 15 – 15, 20 – 20 – 20 – 20, 15 – 20 – 20 – 20, and 10 – 30 – 30 – 30 panels, respectively, under the small-difficulty difference condition. The four plots in the second column present the same information derived under the large-difficulty difference condition. The bias plot displays the theta differences (e.g., estimated minus true) associated with each proficiency estimator. In Figure 5, the four plots in the first column present estimation error derived from the four module-length conditions, respectively, under the small-difficulty difference condition. The second column presents the same information derived under the large-difficulty difference. The error plot displays conditional standard deviations of the differences, which can be interpreted as empirical estimates of the conditional standard errors of measurement. In Figure 6, the four plots in the first column present RMSEs derived from the four module-length conditions, respectively, under the small-difficulty difference condition. The second column presents the same information derived under the large-difficulty

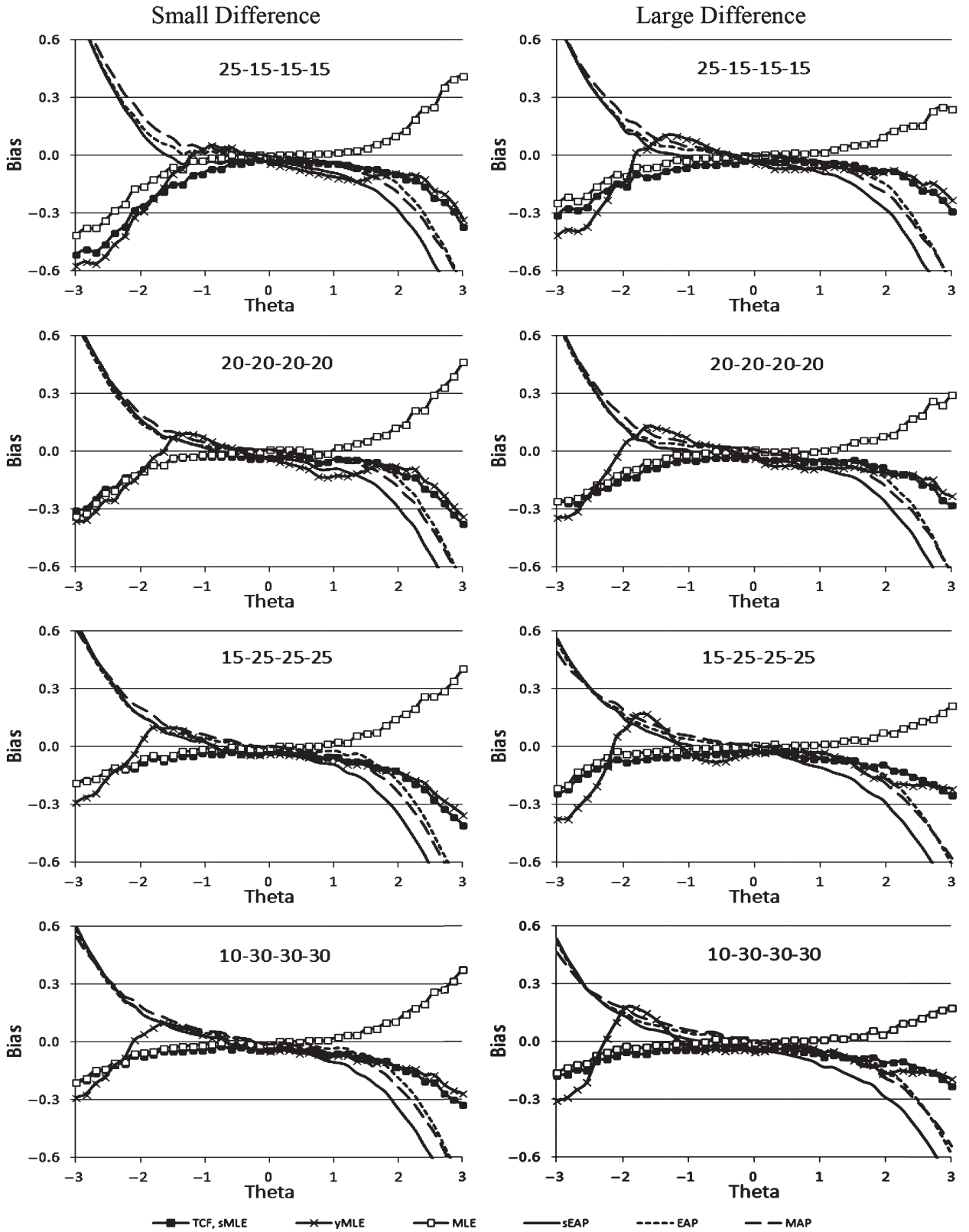

Figure 4 Conditional bias of each estimation method under the small- and large-difference conditions. TCF = test characteristic function; $_{S}$MLE = maximum likelihood estimation with summed scoring; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori with item-pattern scoring; MAP = maximum (mode) a posterior.

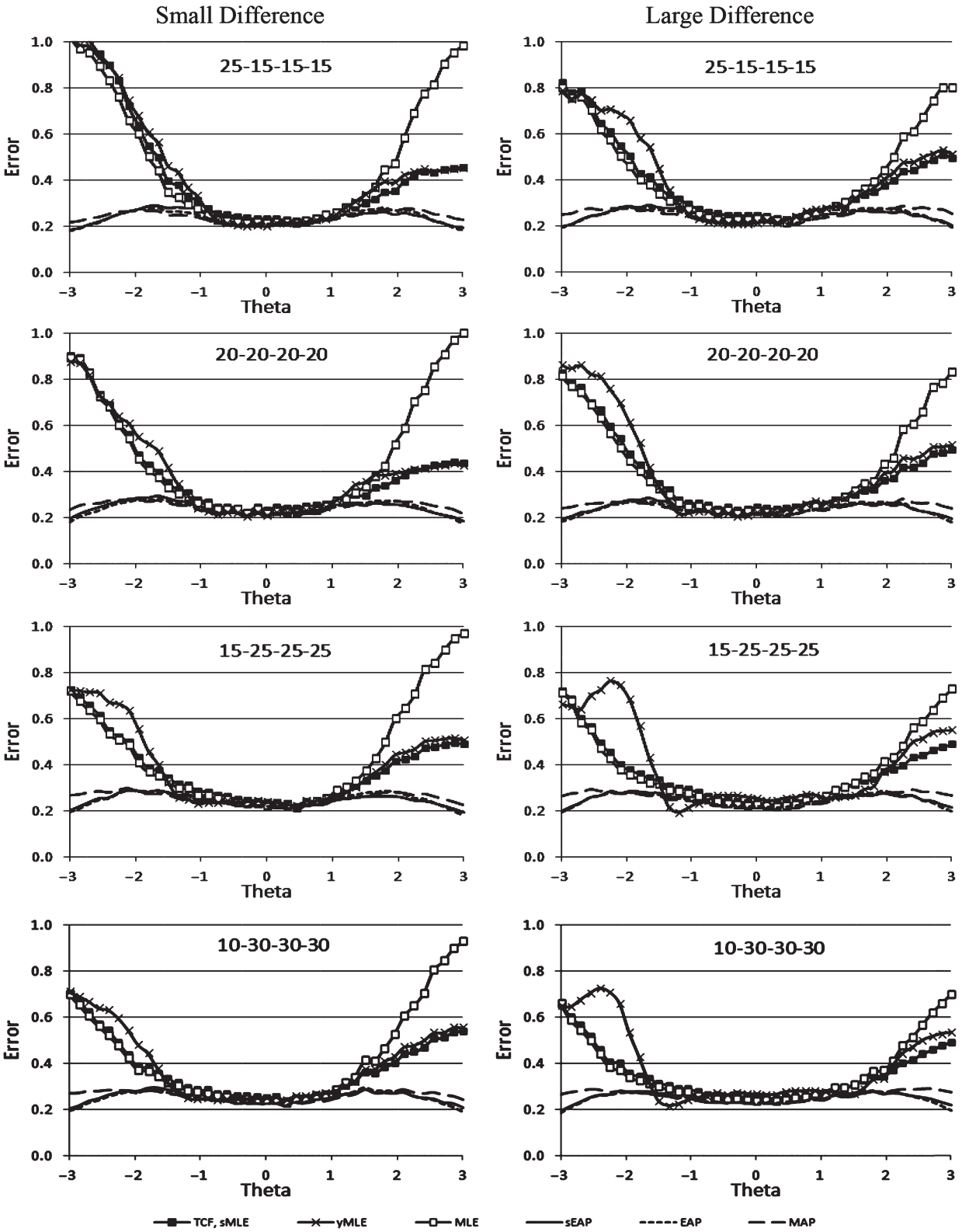

Figure 5 Conditional error of each estimation method under the small- and large-difference conditions. TCF = test characteristic function; $_{S}$MLE = maximum likelihood estimation with summed scoring; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori with item-pattern scoring; MAP = maximum (mode) a posterior.

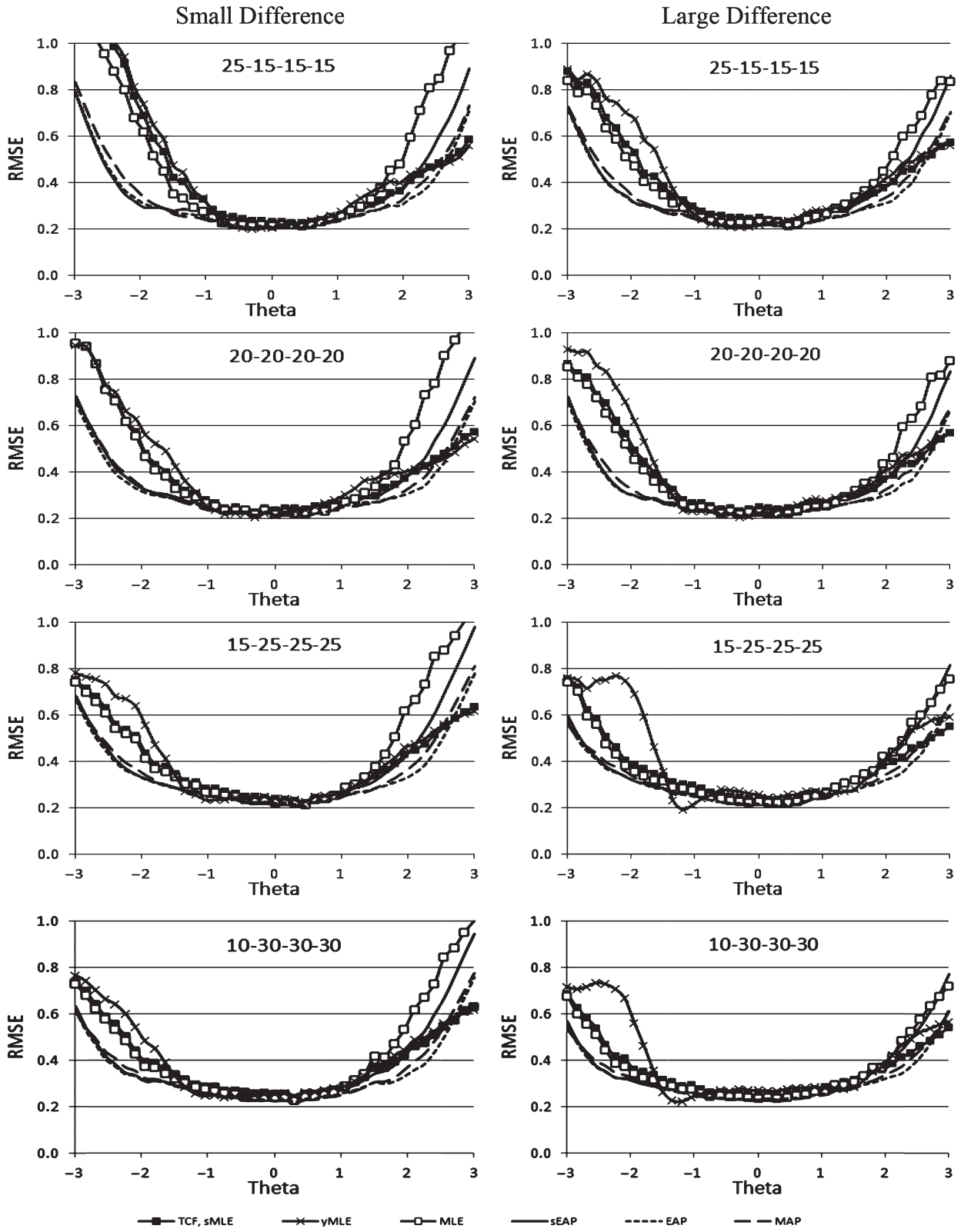

Figure 6 Conditional root-mean-square error of each estimation method under the small- and large-difference conditions. TCF = test characteristic function; $_{S}$MLE = maximum likelihood estimation with summed scoring; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori with item-pattern scoring; MAP = maximum (mode) a posterior.

Table 2 Summary of Deviance Measures for Each Estimation Method Under Small- and Large-Difficulty Difference Conditions

		Small difference			Large difference		
MST panel	Method	Bias	Error	RMSE	Bias	Error	RMSE
25–15–15–15	TCF	0.093	0.304	0.318	0.061	0.290	0.297
	$_{Y}$MLE	0.104	0.308	0.325	0.075	0.294	0.304
	$_{S}$MLE	0.095	0.305	0.319	0.062	0.291	0.297
	MLE	0.061	0.303	0.309	0.042	0.285	0.288
	$_{S}$EAP	0.110	0.240	0.264	0.107	0.244	0.266
	EAP	0.078	0.230	0.243	0.077	0.234	0.246
	MAP	0.095	0.225	0.244	0.091	0.230	0.247
20–20–20–20	TCF	0.055	0.286	0.291	0.061	0.282	0.289
	$_{Y}$MLE	0.084	0.287	0.299	0.076	0.283	0.293
	$_{S}$MLE	0.055	0.287	0.292	0.062	0.283	0.289
	MLE	0.052	0.295	0.300	0.042	0.279	0.282
	$_{S}$EAP	0.111	0.242	0.266	0.102	0.240	0.261
	EAP	0.075	0.235	0.246	0.072	0.231	0.242
	MAP	0.090	0.231	0.247	0.085	0.228	0.243
15–25–25–25	TCF	0.062	0.284	0.291	0.055	0.277	0.282
	$_{Y}$MLE	0.067	0.287	0.294	0.076	0.294	0.304
	$_{S}$MLE	0.062	0.285	0.292	0.055	0.277	0.282
	MLE	0.048	0.299	0.303	0.026	0.271	0.273
	$_{S}$EAP	0.123	0.244	0.273	0.110	0.243	0.267
	EAP	0.082	0.239	0.253	0.074	0.236	0.247
	MAP	0.098	0.235	0.254	0.085	0.233	0.248
10–30–30–30	TCF	0.059	0.290	0.296	0.055	0.279	0.285
	$_{Y}$MLE	0.065	0.292	0.299	0.068	0.292	0.300
	$_{S}$MLE	0.059	0.291	0.296	0.055	0.279	0.285
	MLE	0.042	0.298	0.301	0.023	0.274	0.275
	$_{S}$EAP	0.122	0.251	0.279	0.107	0.249	0.271
	EAP	0.081	0.246	0.259	0.071	0.243	0.253
	MAP	0.094	0.241	0.259	0.081	0.241	0.254

Note. TCF = test characteristic function; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$MLE = maximum likelihood estimation with summed scoring; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori with item-pattern scoring; MAP = maximum (mode) a posterior.

difference. The RMSE plot displays the conditional total error, which is a combination of bias and error. Only six lines appear in all plots, because the difference between TCF and $_{S}$MLE is almost indistinguishable with respect to all deviance measures. Accordingly, any explanations related to TCF apply directly to $_{S}$MLE as well. Table 2 presents a summary of the weighted deviance measures, averaged across the entire theta scale from −3.0 to +3.0, under the small- and large-difficulty difference conditions.

Under the small-difficulty difference condition, as shown in Figure 4, the seven estimators' magnitudes of bias were nearly identical across the theta region from −1.0 to +1.0. The magnitude of bias associated with the seven estimators differed noticeably in the top and bottom theta regions. The same trend appeared across the four MST panel types. At the extremes, MLE produced the least bias, TCF/$_{S}$MLE and $_{Y}$MLE produced the second least bias, and the Bayesian methods produced the largest bias. Among the three Bayesian methods, $_{S}$EAP produced larger biases than did EAP or MAP in the upper region of the theta scale. As presented in Table 2, the average weighted RMSEs of the three Bayesian methods were larger than those of MLE and TCF/$_{S}$MLE. Four non-Bayesian methods produced smaller bias at the extremes of the theta scale using the 15–25–25–25 and 10–30–30–30 panels, in which more items appeared in the second-stage module.

For conditional error under the small-difficulty difference condition, all methods performed similarly across the theta region from −1.5 to +1.5. At the extremes of the theta scale, three Bayesian methods produced much less error than did TCF/$_{S}$MLE, $_{Y}$MLE, and MLE. MLE yielded the largest error in both extreme regions, whereas TCF/$_{S}$MLE yielded much smaller error in the upper theta region than did MLE. At the lower end of the theta scale, both TCF/$_{S}$MLE and MLE produced smaller error using the 15–25–25–25 and 10–30–30–30 panels. In addition, the difference between TCF/$_{S}$MLE and $_{Y}$MLE was somewhat noticeable under those conditions. The same trends did not emerge, however, in the upper region. The pattern of the three Bayesian methods was very consistent across the four MST panels.

As a consequence of the reduction in standard error, three Bayesian methods yielded smaller RMSEs than did their non-Bayesian counterparts, mainly in the top and bottom theta regions. The difference between the non-Bayesian methods and their Bayesian counterparts was rather salient in a situation where more items appeared at Stage 1 (e.g., 25 – 15 – 15 – 15). Under the small-difficulty difference condition, as shown in Figure 6, overall error (i.e., RMSE) of the seven estimators was very similar across the theta region from −1.5 to +1.5. As the average weighted RMSEs indicate, the overall RMSEs of the Bayesian methods were slightly smaller than those of the non-Bayesian methods. Among the three Bayesian methods, $_S$EAP produced larger RMSEs than did EAP or MAP mainly in the upper theta region. MLE produced much larger RMSEs at the top of the theta scale than did TCF/$_S$MLE/$_Y$MLE. $_Y$MLE produced the largest RMSEs at the bottom of the theta scale using the 15 – 25 – 25 – 25 and 10 – 30 – 30 – 30 panels. The seven estimators' relative performances were very consistent for the four MST panels. Even so, the difference between the non-Bayesian methods and their Bayesian counterparts substantially decreased using the 10 – 30 – 30 – 30 panel. On average, as shown in Table 2, all three Bayesian methods produced the smallest RMSEs using the 25 – 15 – 15 – 15 panels, whereas the non-Bayesian methods produced the largest RMSEs using the same panel.

The large-difference in difficulty condition displayed smaller bias, error, and RMSE than did the small-difference condition, particularly in the two extreme theta regions, because the second-stage modules under the large-difference condition achieved a better measurement precision in those regions. Consequently, as shown in Table 2, the average weighted RMSEs were generally smaller under the large-difference condition. This trend was replicated across most panel types. $_Y$MLE performed somewhat differently, however, leading to slightly larger error and RMSE using the 15 – 25 – 25 – 25 and 10 – 30 – 30 – 30 panels under the large-difference condition. In general, using distinct second-stage modules was beneficial for most estimators, primarily MLE. The distinction between the small- and large-difference conditions was salient in the assembly condition in which a large number of items appeared in the second stage (e.g., 10 – 30 – 30 – 30). The difference between Bayesian and non-Bayesian methods was relatively smaller under the large-difference 10 – 30 – 30 – 30 panel assembly condition than under the other assembly conditions.

As an auxiliary analysis, we used the uniform distribution as a prior for the three Bayesian ($_S$EAP, EAP, and MAP) estimators to evaluate their sensitivity to the choice of a prior distribution. The uniform prior limits the range of allowable theta (θ) values and avoids the issue of infinite estimates of proficiency (Magis et al., 2011). Regardless of prior, relative performance of the six proficiency estimators, except for MAP,[6] was almost identical across the eight MST panel assembly conditions.[7]

Figure 7 presents percentage correct classifications of the seven estimators at Stage 1 (routing). The four plots represent the four module-length conditions, respectively. Because the same routing module was used in both the small- and large-difficulty difference conditions, the percentage correct classification result for each method was essentially the same in both difference conditions. Each plot has six lines (TCF = $_S$MLE) that represent the percentage correct classification of each of the seven estimators across the theta region where most misclassifications take place, from −1.5 to +1.5. Table 3 presents the percentage correct classification of each method in a particular theta region, along with its average weighted percentage correct classification.

In general, all seven methods produced similar patterns. The percentage correct classification differed as a function of the panel design. As expected, all methods generally performed better with the 25 – 15 – 15 – 15 panel, compared to the 10 – 30 – 30 – 30 panel, where only 10 items were available to estimate proficiency levels for routing. As summarized in Table 3, the average differences between the two panels were slightly more than 10%. Number-correct scoring yielded slightly higher classification error than did item-pattern scoring at the middle theta region (around −0.75 to +0.75) using the 15 – 25 – 25 – 25 panel. Compared to other methods, EAP and MAP produced better classifications using the 10 – 30 – 30 – 30 panel. Non-Bayesian methods seem more sensitive to a lack of resources than Bayesian methods, leading to a lower percentage of correct classifications. As expected, the correct classification rates were lower at the regions where the two cutscores (−0.5 and +0.5) were located, because misclassification more likely occurs for borderline examinees whose proficiency levels are close to the cutscores.

Discussion

Using simulated data sets, we investigated the effectiveness of IRT proficiency estimators under various two-stage MST panel assembly conditions. Using real data sets (from linear tests), Kolen and Tong (2010) showed that the choice of

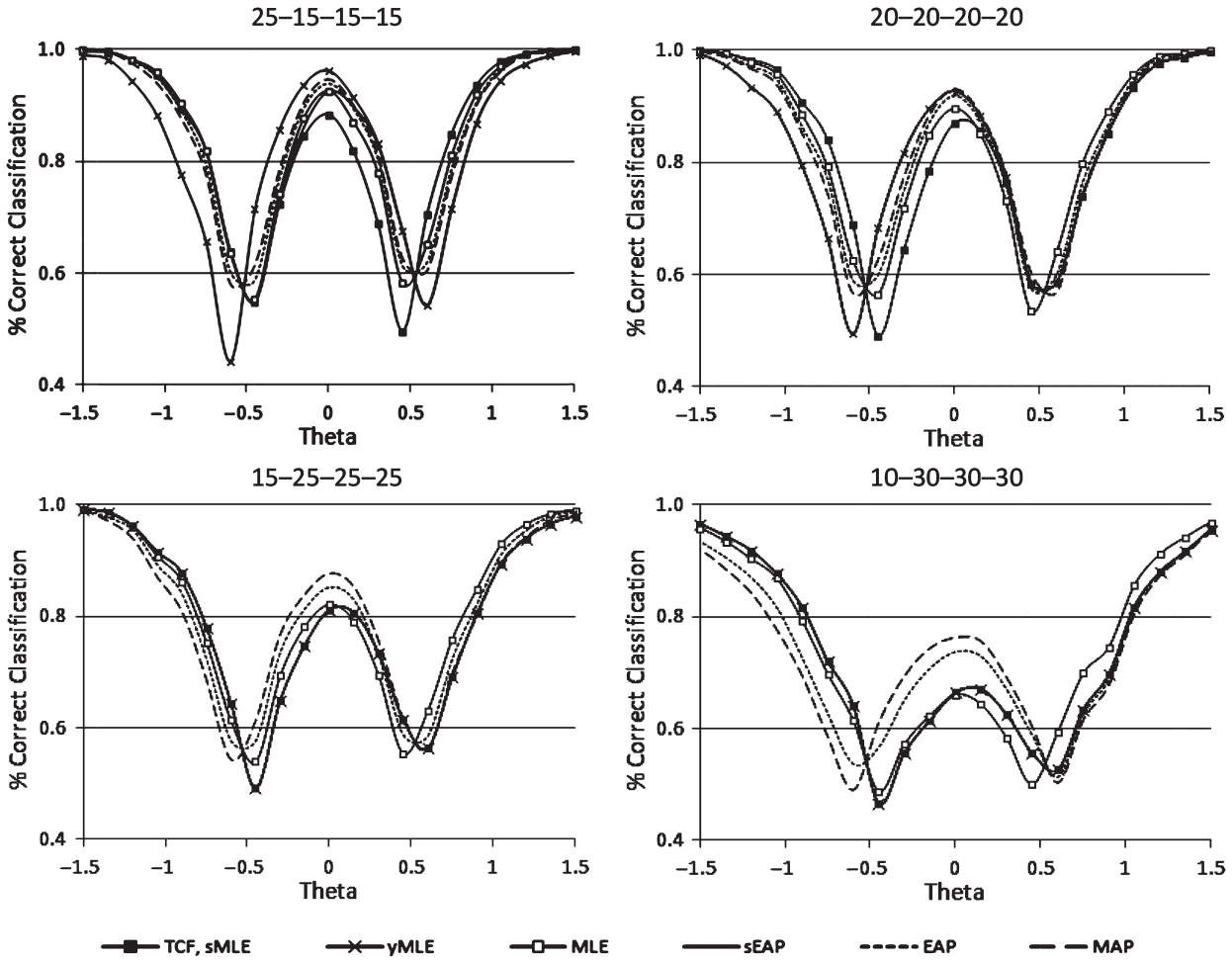

Figure 7 Percentage correct classification of each estimation method at Stage 1. TCF = test characteristic function; $_{S}$MLE = maximum likelihood estimation with summed scoring; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; MAP = maximum (mode) a posterior.

Bayesian (prior) versus non-Bayesian (no prior) estimators was of more practical significance than the choice of number-correct versus item-pattern scoring. This trend was replicated in the present simulation study. The magnitude of deviances such as bias and error varied depending on the choice of IRT proficiency estimators, mainly at the top and bottom regions of the theta scale. Bayesian estimators using prior information yielded greater bias but smaller standard errors than did their non-Bayesian counterparts. For the extreme proficiency levels ($\theta < -1.5$ or $1.5 < \theta$), the decrease in standard error compensated for the bias characteristic of Bayesian estimates, resulting in smaller overall error (i.e., RMSE). As expected, use of prior information will be a promising way to assign more accurate proficiency levels to these examinees. We concluded that possible score changes caused by the use of different proficiency estimators will be nonnegligible, particularly for low- and high-performing examinees.

The choice of a prior distribution of proficiency levels is a key factor associated with Bayesian estimators' effectiveness. The mean proficiency of Bayesian estimators varied as a function of the choice of prior distributions. According to Kolen and Tong (2010), the use of a prior distribution with a higher mean tended to increase the percentage at the highest proficiency level but decrease the percentage at the lowest level. Conversely, the use of a prior distribution with a lower mean led to the opposite effect. Applying a standard normal prior distribution, Bayesian estimators' bias decreases with theta (Bias $\left(\widehat{\theta}\right) = \left(\widehat{\theta} - \theta\right)$) and is approximately equal to zero for proficiency levels close to the prior mean value. Owing to the shrinkage phenomenon associated with the use of priors, Bayesian estimators overestimate negative proficiency levels but underestimate positive ones. Thus the magnitude of bias was large at the extreme levels. As expected, both the EAP and MAP estimators yielded close estimates of proficiency levels.

Table 3 Correct Classification Percentages of Each Proficiency Estimation Method at Stage 1

Panel type	Theta range	TCF	$_{Y}$MLE	$_{S}$MLE	MLE	$_{S}$EAP	EAP	MAP
25 – 15 – 15 – 15	−3.00 to −0.90	99	97	99	99	99	99	99
	−0.75 to −0.30	68	67	68	69	68	69	69
	−0.15 to 0.15	85	94	85	89	89	91	92
	0.30 to 0.75	68	69	68	71	69	70	70
	0.90 to 3.00	99	98	99	99	98	99	99
	Weighted average	83	83	83	84	83	85	85
20 – 20 – 20 – 20	−3.00 to −0.90	99	97	99	99	97	98	98
	−0.75 to −0.30	67	67	67	68	67	68	68
	−0.15 to 0.15	84	90	84	86	90	89	90
	0.30 to 0.75	67	67	67	68	67	67	67
	0.90 to 3.00	98	98	98	99	98	99	98
	Weighted average	82	82	82	83	82	83	83
15 – 25 – 25 – 25	−3.00 to −0.90	98	98	98	98	98	98	97
	−0.75 to −0.30	64	64	64	65	64	65	65
	−0.15 to 0.15	79	79	79	80	79	83	86
	0.30 to 0.75	65	65	65	66	65	66	66
	0.90 to 3.00	97	97	97	98	97	98	97
	Weighted average	79	79	79	80	79	80	80
10 – 30 – 30 – 30	−3.00 to −0.90	97	97	97	96	97	94	93
	−0.75 to −0.30	60	60	60	59	60	60	59
	−0.15 to 0.15	65	65	65	64	65	72	75
	0.30 to 0.75	59	59	59	59	59	60	60
	0.90 to 3.00	95	95	95	96	95	95	94
	Weighted average	72	72	72	73	72	73	73

Note. TCF = test characteristic function; $_{Y}$MLE = Yen's (1984) maximum likelihood estimation; $_{S}$MLE = maximum likelihood estimation with summed scoring; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori with item-pattern scoring; MAP = maximum (mode) a posterior.

The conditional errors of the estimators were reflective of two major issues, including the use of item-pattern scoring or number-correct scoring in the estimation and the shrinkage that occurs with the use of prior information in the Bayesian estimators. As described elsewhere (Kolen & Tong, 2010; Lord, 1980), the finding that the conditional variance of the TCF estimator is larger than that of the MLE estimator reflects the greater error and greater loss of information from the use of number-correct scores in TCF. Somewhat different from Kolen and Tong's (2010) findings, the current study found that the conditional variance of the $_{S}$EAP estimator was greater than the variance of the EAP and MAP estimators. The differences in the estimators appear to be due to whether the variances being compared are conditional (this study) or unconditional (Kolen & Tong, 2010, Table 1) and also whether the estimation is based on item response patterns (EAP) or number-correct scores ($_{S}$EAP). That is, number-correct scores used in the $_{S}$EAP estimator are associated with greater conditional error and are subjected to less shrinkage in the $_{S}$EAP estimation than the item response patterns used in EAP. These issues play out differently in overall errors because overall errors are functions not only of conditional errors but also of the variances of the conditional means, which reflect the different shrinkages of the EAP and $_{S}$EAP estimators. The differences in how shrinkage affects the conditional and overall errors of the estimators account for the differences in the ordering of conditional errors ($_{S}$EAP > EAP) in our study and overall errors (EAP > $_{S}$EAP) in Kolen and Tong (2010, Table 1; C. Lewis, personal communication, November 6, 2013).

We compared two versions of number-correct scoring MLE, $_{Y}$MLE and $_{S}$MLE. The revised version ($_{S}$MLE) performed slightly better than did the original version ($_{Y}$MLE), leading to smaller deviance measures in most panel design conditions. We concluded that the exact recursion algorithm will be more effective in estimating proficiency level based on the likelihood for the number-correct score than the approximation method proposed by Yen (1984). In addition to that, the result of $_{S}$MLE was indistinguishable from the result of TCF across the entire theta region (except for the extremes) in all the panel design conditions. TCF generally uses the golden section search method by plugging successive guesses into the TCC function until the result is the observed number-correct score. This study confirmed that TCF will be as effective as the MLE approach as long as number-correct scoring is used as a source of the likelihood function. Although the idea of the number-correct scoring MLE is intuitively appealing, its psychometric benefit may be negligible.

Certainly the two versions of number-correct scoring MLE employed in this study act more like TCF, not like item-pattern scoring MLE.

The large-difference conditions led to smaller bias, error, and RMSE than did the small-difference conditions, particularly at the low and high regions of the theta scale. The difference between the small and large conditions was crucial when a large number of items appeared in the second stage (e.g., 10 – 30 – 30 – 30). This finding indicates that the performance of proficiency estimators could be associated with the measurement characteristics of Stage 2 modules. The comparison between the two MST assembly conditions, small- versus large-difficulty differences, shows a benefit from the use of distinct Stage 2 modules under a two-stage MST. In particular, the RMSE of the seven estimators was relatively similar in size across the most score scale (−2.0 to +2.0) when many items distinct in difficulty were administered (e.g., 10 – 30 – 30 – 30). In this panel condition, many more items in the difficult module targeted strong performers, whereas many more items in the easy module targeted weak performers. Maintenance of similar levels of measurement precision across the entire score scale is promising for assessments in which all score points are equally important. In particular, testing programs that offer special benefits (e.g., scholarships) to outstanding performers must produce precise and accurate proficiency estimates at the extreme upper range of the scale. In that case, Bayesian methods may be a good option for practical use. To enhance estimation accuracy, the choice of a prior distribution could be considered. One might choose either a normal distribution or a uniform distribution as a prior, as in this simulation study. Even a score distribution derived from a sufficiently large amount of operational data could be used as a prior.

Concerns have been raised about the potential score variability caused by routing error, particularly for borderline examinees whose scores are near the routing cutscores. In reality, however, those borderline examinees' scores would not be heavily influenced by the choice of subsequent-stage module. Perhaps potential differences in reported scores caused by routing itself are minimal. More dramatic changes would appear if clearly strong or clearly weak performers received a module that was not best matched with their actual proficiency levels.

We also compared the performance of the seven estimators in terms of the percentage correct classification at Stage 1. As expected, the more items in the routing module there are, the more accurate the assignment to the second-stage module will be. A 20-item routing module performed as well as a 25-item routing module, whereas 10- and 15-item routing modules performed poorly in the theta region where routing cutoff scores are located. Even so, the proficiency accuracy for each of the seven estimators was generally similar across the four module-length conditions for most examinees. This finding implies that the impact of misrouting would be minimal under the MST design employed in this study. Although the magnitude of misrouting was substantial with the 10 – 30 – 30 – 30 panel, its final proficiency estimates were as accurate as those in other panels. H. Kim and Plake (1993) found that the statistical characteristics of the first-stage (routing) module had a major influence on the complete test's measurement precision compared to the later stages' modules. This may not be true, however, when a sufficient number of items is administered at the subsequent stage. This means that the number of items at Stage 1 would not be more important than the numbers of items at subsequent stages. By design, proficiency estimates for both low- and high-performing examinees were more accurate under the assembly designs in which more items appeared at the later stage. In reality, however, most examinees would be located at the theta region from −1.5 to +1.5; thus any module-length combinations under a two-stage MST would have similar effectiveness in estimating proficiency levels. Estimates of examinees' proficiency may be robust enough to shield them from the effects of misrouting, particularly in the two-stage MST design used in this study, unless MST modules include many ill-fitting and unreliable items. This finding is compatible with a previous study (S. Kim & Moses, 2014) that showed the minimal impact of misrouting under a two-stage MST scored by TCF.

On the basis of the findings, we concluded that the impact of IRT proficiency estimators would be minimal as long as a sufficient number of high-quality items is available with which to estimate the examinees' proficiency levels. In particular, under the simplest MST design framework (e.g., two stages, three paths), some variations imposed on the MST panel assembly may not produce a substantial impact on the accuracy of proficiency estimates as long as the same number of items is used to estimate the proficiency levels. If all outcomes are identical, a simple scoring method (e.g., TCF) and a simple MST panel assembly may be a good choice for practical purposes. If the item pool size is small because of limited resources, however, the difference among the proficiency estimators might be nontrivial, not only at the extreme regions, but also in the middle region of the theta scale. The use of items distinct in difficulty may lead to more accurate proficiency estimates under MST, but it is often demanding for item writers to create well-functioning difficult items in reality.

There are limitations in generalizing the findings of the current study in practice. First, we assembled eight MST panels to meet the two assembly conditions (2 difficulty differences × 4 module lengths). In a particular assembly condition, a single particular panel was used to generate all examinees' responses. Therefore any investigation between IRT estimators was straightforward in this application. Even so, the use of numerous parallel MST panels in each simulation should be close to the real testing environment. It is premature to determine which panel design will be the best under the two-stage MST based solely on the present findings. Second, we compared the proficiency estimates of each estimator to the "true" values at the theta scale. In a real testing setting, however, testing programs report scaled scores to the examinees after applying a particular transformation procedure to the thetas to make them comparable over time. Therefore, the impact of any differences among the IRT proficiency estimators on the reporting scale would be of practical significance, because that has a direct impact on reported scores. Further investigation using a hypothetical score scale is ongoing.

Additional studies are needed to achieve solid conclusions about MST design and its practical implications. In this study, we compared the performance of various IRT proficiency estimators using the MST panels that contained well-fitting items in discrimination. We considered the second-stage modules' difference only in difficulty, not in discrimination. It would be worthwhile to consider item discrimination factors as well to see if using more difficult (or easier) but less discriminating items would confirm the current findings. Particularly, comparing number-correct scoring with item-pattern scoring would be interesting in a situation in which misrouting takes place but poor-quality items are included in subsequent-stage modules. We used the simulated data to obtain clear conclusions regarding the IRT estimators' accuracy. Among the eight panel conditions employed in this study, a particular panel perfectly reflects actual panels that had been administered under operational testing settings. Even so, comparisons among the IRT estimation methods using real data sets would be informative.

Many studies examining MST tend to focus on design issues such as adequate numbers of stages, numbers of modules, or panel assembly (see Jodoin et al., 2006; Luecht & Nungester, 1998; Luecht et al., 1996; Wang et al., 2012). Very little guidance appears in the literature about the effects of the choice of an estimator in practical applications. This is a very important topic, because the choice of scoring method has a direct influence on the examinees' reported scores or proficiency levels. We think that the present findings will provide some realistic guidelines to practitioners who want to adopt MST procedures in their assessments. A short note of this paper can be found elsewhere (Kim, Moses, & Yoo, 2015).

Notes

1 Number-correct scoring is interchangeable with summed scoring.
2 Warm (1989) proposed the weighted likelihood estimator (WLE), whose estimation values are based on the mean of the likelihood function. Consequently, WLE estimates are generally slightly more central than MLE estimates, but the difference between MLE and WLE would be small.
3 The 2PL model is used operationally for some large-scale testing programs (e.g., the *GRE*® revised General test and the *TOEFL*® test).
4 The range of values for the estimated proficiency level is usually restricted so that the maximum likelihood method provides finite estimates. We set an upper bound (+5.0) for the only-correct score and a lower bound (−5.0) for the only-incorrect score.
5 We simulated the examinees' theta distribution to be uniform to obtain stable estimates over the entire theta scale. To compute the summary statistics, however, we applied the normal distribution weights to reflect a more realistic distribution of examinees' thetas.
6 The range of values for the estimated proficiency level is usually restricted so that the maximum likelihood method provides finite estimates. Applying such a range restriction is equivalent to estimating the proficiency level with a Bayesian estimator and a uniform uninformative prior distribution on the selected range (Magis et al., 2011). If the prior distribution is uniform (i.e., uninformative prior), then the Bayesian modal estimator (i.e., MAP) is essentially identical to the MLE (Yen & Fitzpatrick, 2006).
7 For simplicity, the estimation results derived from the uniform prior case were not presented in the summary tables and figures. The authors can provide the results on request.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.
Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431 – 444.

Educational Testing Service. (2011). *GRE information and registration bulletin*. Princeton, NJ: Author.

Green, D. R., & Yen, W. M. (1983, April). *Number-correct versus pattern scoring*: *Results for ethnic groups*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203 – 220.

Kim, S., & Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing*: *A simulation study* (Research Report No. RR-14-01). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12000

Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*, 70 – 79.

Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*, 8 – 14.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *8*, 453 – 461.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet-assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*, 189 – 202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229 – 249.

Luecht, R. M., Nungester, R. J., & Hadidi, A. (1996, April). *Heuristic-based CAT*: *Balancing item information*, *content and exposure*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing* (Research Report No. 2011 – 12). New York, NY: College Board.

Magis, D., Beland, S., & Raiche, G. (2011). A test-length correction to the estimation of extreme proficiency levels. *Applied Psychological Measurement*, *35*, 91 – 109.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201 – 210.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227 – 253.

Tong, Y., & Kolen, M. J. (2010, April). *IRT proficiency estimators and their impact*. Paper presented at the annual meeting of the National Council in Measurement in Education, Denver, CO.

Wang, X., Fluegge, L., & Luecht, R. (2012, April). *A large-scale comparative study of the accuracy and efficiency of ca-MST*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, *54*, 427 – 450.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, *21*, 93 – 111.

Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education*, *4*(3), 209 – 228.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111 – 153). Westport, CT: Praeger.

Suggested citation:

Kim, S., Moses, T., & Yoo, H. (2015). *Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing* (Research Report No. RR-15-11). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12057

Action Editor: Marna Golub-Smith

Reviewers: Longjuan Liang and HongwenGuo